FEM: Basic Theory and Implementation

1-D FEM for Elliptic Equation

Huaijin Wang

March 18, 2025

Contents

1 Introduction

- Why Elliptic problem?
- Why 1D case?
- Why FEM?

2 Elliptic Problem

We consider the elliptic problem:

2.1 Typical Model: Possion Equation with Homogeneous Dirichlet Boundary

A two-point boundary value problem with homogeneous Dirichlet boundary condition:

$$\begin{cases} -\dfrac{d^2u}{dx^2} = f(x), \quad x \in I := (0,1), \\ u(0) = u(1) = 0, \end{cases} \tag{2.1}$$

where $f \in L^2(I)$. The problem (2.1) is also called the *strong problem*. Let $V := H_0^1(I) = \{v \in H^1(I) : v(0) = v(1) = 0\}$. The restriction on boundary values makes sense due to the embedding theorem, which tells that

$$\forall v \in H^1(I),\ \exists \bar{v} \in C(\bar{I}) \text{ s.t. } v = \bar{v} \text{ a.e. in } I.$$

The *variational problem* (or known as *weak problem*):

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ (u', v') = (f, v), \quad \forall v \in V, \end{cases} \tag{2.2}$$

where $(\cdot,\cdot)$ stands for the $L^2(I)-$inner product. Let J be the linear functional:

$$J(v) = \frac{1}{2}(u', v') - (f, v).$$

Then the *minimization problem*:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ J(u) \leqslant J(v), \quad \forall v \in V. \end{cases} \tag{2.3}$$

The term minimization problem corresponds the "principle of minimum potential energy" in mechanics. It tells us that some of differential equations like (2.1) may originates from minimizing the potential energy in some physical problems.

Theorem 2.1. *Under proper regularity assumptions, the three problems above are equivalent:*
1). the solution of (2.1) *is a solution of* (2.2)*;*
2). the solution of (2.2) *is a solution of* (2.3)*;*
3). the solution of (2.3) *is a solution of* (2.1).

The existence and uniqueness of these three problem can be considered respectively.

For (2.1), its existence can be represented using Green's function (see [Evans (2010), p.35 Chapter 2, Theorem 12]), and the uniqueness is guaranteed by the *strong maximum principle* of harmonic functions (see [Evans (2010), pp. 27-28, Theorem 4&5]).

For (2.3), its existence and uniqueness are guaranteed by that J is strongly convex and is a linear functional over a linear space.

For (2.2), its existence and uniqueness are guaranteed by the well known *Lax-Milgram Lemma*, whose general description reads

Lemma 2.1 (Lax-Milgram)**.** *Let V be a Hilbert space, endowed with the norm $\|\cdot\|_V$. Consider the problem:* $\forall f \in V'$,

$$\begin{cases} \textit{Find } u \in V, \ \textit{such that} \\ a(u,v) =< f, v >, \quad \forall v \in V, \end{cases}$$

where $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a bilinear form. If furthermore, $a(\cdot,\cdot)$ satisfies

$$\text{Continuity}: \quad \exists \gamma > 0 \ s.t. \ |a(u,v)| \leqslant \gamma \|u\|_V \|v\|_V, \quad \forall u, v \in V,$$
$$\text{Coercivity}: \quad \exists \alpha > 0 \ s.t. \ a(v,v) \geqslant \alpha \|v\|_V^2, \quad \forall v \in V.$$

Then the problem admits a unique solution u, which satisfies

$$\|u\|_V \leqslant \frac{1}{\alpha} \sup_{v \in V, v \neq 0} \frac{< f, v >}{\|v\|_V}.$$

Theorem 2.2. *Problem* (2.2) *admits an unique solution.*

2.2 Other Boundary Conditions

3 $P1-$ FEM

The finite element method (FEM) is a numerical technique, arguably the most robust and popular, for solving differential equations. FEM is a numerical method general based on the *Galerkin approximation* (or *Galerkin method* or *Galerkin framework*), to approximate with constructing finite elements (piecewise approximation). Galerkin method is to approximate the weak problem with finite dimensional subspace constructed. For (2.2),

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ \left(\dfrac{du_h}{dx}, \dfrac{dv_h}{dx} \right) = (f, v_h), \quad \forall v_h \in V_h, \end{cases} \tag{3.1}$$

where V_h is a finite dimensional subspace of V.

We divide the interval $[0,1]$ into $N+2$ grid

$$0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1.$$

We denote the subintervals $I_j = [x_{j-1}, x_j]$ for $1 \leqslant j \leqslant N+1$, with length $h_j = x_j - x_{j-1}$. Let $h = \max_{1 \leqslant j \leqslant N+1} h_j$. The mesh size h is used to measure how fine the partition is.

We define the finite element space

$$V_h = \{ v \in C[0,1] : v \text{ is linear on each subinterval } I_j, \text{ and } v(0) = v(1) = 0 \}.$$

Theorem 3.1. *$V_h \subset V$.*

Proof. It is sufficient to show that for any $v \in V_h$ we have $v \in H^1(I)$, i.e.,

$$\int_0^1 \frac{dv}{dx} \phi \mathrm{d}x = -\int_0^1 v \frac{d\phi}{dx} \mathrm{d}x, \quad \forall \phi \in C_0^\infty(I).$$

In fact,

$$\int_0^1 \frac{dv}{dx} \phi \mathrm{d}x = \sum_{j=1}^{N+1} \int_{I_j} \frac{dv}{dx} \phi \mathrm{d}x = \sum_{j=1}^{N+1} \left(\phi(x_j) v(x_j) - \phi(x_{j-1}) v(x_{j-1}) - \int_{I_j} v \frac{d\phi}{dx} \mathrm{d}x \right)$$
$$= \phi(1) v(1) - \phi(0) v(0) - \sum_{j=1}^{N+1} \int_{I_j} v \frac{d\phi}{dx} \mathrm{d}x = -\int_0^1 v \frac{d\phi}{dx} \mathrm{d}x.$$

□

Theorem 3.2. $\dim(V_h) = N$.

Proof. For any $v_h \in V_h$, we observe that on each subinterval I_j for $j = 1, \cdots, N+1$, $v|_{I_j}$ is a linear polynomial and thus uniquely determined by 2 parameters, known as the *degree of freedom*. Since there are $N+1$ subintervals, this initially gives a total of $2(N+1)$ degrees of freedom. However, imposing N continuity conditions at the subinterval boundaries and 2 boundary conditions reduces the count by $N+2$, leaving $2(N+1) - N - 2 = N$ degrees of freedom. Consequently, the dimension of the space is N. $\square$

Remark 3.1. *Why nodal basis functions?*

Let us introduce the linear basis function $\phi_j(x)$ for $1 \leqslant j \leqslant N$, which satisfies the properties

$$\phi_j(x_i) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

Then $\phi_j(x) \in V_h$ and $\{\phi_1(x), \cdots, \phi_N(x)\}$ is linear independent and thus, by dimension argument, constitutes a basis for V_h, i.e., $V_h = \operatorname{span}\{\phi_1, \cdots, \phi_N\}$. Consequently, $\forall v_h \in V_h$, there is an unique representation

$$v_h(x) = \sum_{j=1}^{N} v_j \phi_j(x), \quad x \in [0,1],$$

where $v_j = v_h(x_j)$. More specifically, ϕ_j is given by

$$\phi_j(x) = \begin{cases} \frac{x - x_{j-1}}{h_j}, & \text{if } x \in [x_{j-1}, x_j], \\ \frac{x_{j+1} - x}{h_{j+1}}, & \text{if } x \in [x_j, x_{j+1}], \\ 0, & \text{elsewhere}. \end{cases} \tag{3.2}$$

With the constructed piecewise linear space $V_h = \operatorname{span}\{\phi_1, \cdots, \phi_N\}$, we set the solution u_h of (3.1) as

$$u_h(x) = \sum_{j=1}^{N} u_j \phi_j(x), \quad u_j = u_h(x_j).$$

Substituting u_h in (3.1) and choosing $v = \phi_i(x)$ in (3.1) for each $i = 1, \cdots, N$, we obtain

$$\sum_{j=1}^{N} \left(\frac{d\phi_j}{dx}, \frac{d\phi_i}{dx} \right) u_j = (f, \phi_i) \quad 1 \leq i \leq N,$$

which is a linear system of N equations with N unknowns u_j:

$$\mathbf{A}\mathbf{u} = \mathbf{F},$$

where $\mathbf{u} = [u_1, \cdots, u_N]^{\mathrm{T}}$, $\mathbf{F} = [F_1, \cdots, F_N]^{\mathrm{T}}$ with elements $F_i = (f, \phi_i)$, and $\mathbf{A} = (a_{i,j})$ is an $N \times N$ matrix with elements $a_{i,j} = (\frac{d\phi_j}{dx}, \frac{d\phi_i}{dx})$.

The matrix $\mathbf{A}$ is called the *stiffness matrix* and $\mathbf{F}$ the *load vector*. We can explicitly calculate the elements in $\mathbf{A}$:

$$\begin{aligned}
a_{j,j} &= \left(\frac{d\phi_j}{dx}, \frac{d\phi_j}{dx} \right) = \int_{x_{j-1}}^{x_j} \frac{1}{h_j^2} dx + \int_{x_j}^{x_{j+1}} \frac{1}{h_{j+1}^2} dx = \frac{1}{h_j} + \frac{1}{h_{j+1}}, \quad 1 \leq j \leq N, \\
a_{j-1,j} &= \left(\frac{d\phi_j}{dx}, \frac{d\phi_{j-1}}{dx} \right) = \int_{x_{j-1}}^{x_j} \frac{-1}{h_j^2} dx = -\frac{1}{h_j}, \quad 2 \leq j \leq N, \\
a_{j,j-1} &= \left(\frac{d\phi_{j-1}}{dx}, \frac{d\phi_j}{dx} \right) = a_{j-1,j} = -\frac{1}{h_j}, \quad 2 \leq j \leq N, \\
a_{i,j} &= \left(\frac{d\phi_j}{dx}, \frac{d\phi_i}{dx} \right) = 0, \quad \text{if} \quad |j - i| > 1.
\end{aligned}$$

Thus the matrix $\mathbf{A}$ is tri-diagonal. Let $\mathbf{v} = [v_1, \cdots, v_N]^{\mathrm{T}}$, and we note that

$$\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v} = \sum_{i,j=1}^{N} a_{i,j} v_i v_j = \sum_{i,j=1}^{N} v_j \left(\frac{d\phi_j}{dx}, \frac{d\phi_i}{dx}\right) v_i = \left(\sum_{j=1}^{N} v_j \frac{d\phi_j}{dx}, \sum_{i=1}^{N} v_i \frac{d\phi_i}{dx}\right) = \left(\frac{dv_h}{dx}, \frac{dv_h}{dx}\right) \geqslant 0,$$

where we denote $v_h(x) = \sum_{j=1}^{N} v_j \phi_j(x)$. Thus the equality holds if and only if $\frac{dv_h}{dx} \equiv 0$, which is equivalent to $v_h(x)$ is constant, and by $v_h(0) = 0$ we have $v_h(x) \equiv 0$, or $\mathbf{v} = \mathbf{0}$. Therefore $\mathbf{A}$ is positive definite, which guarantees the linear system has a unique solution.

- $\mathbf{A}$ is symmetric: $a_{i,j} = a_{j,i}$,
- $\mathbf{A}$ is sparse: $a_{i,j} = 0$ for $|i - j| > 1$,
- $\mathbf{A}$ is positive definite.

In a particular case: $h_j = h = \frac{1}{N+1}$, we have

$$\mathbf{A} = \frac{1}{h} \begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{bmatrix}_{N\times N}$$

Theorem 3.3. *Eigenvalue of* $\mathbf{A}$ *is*

3.1 Error Estimate For $P1-$FEM

Let $u \in C(\bar{I})$. We denote u_I the interpolation of u into V_h at nodes $\{x_j\}_{j=0}^{N}$, i.e., $u_I \in V_h$ and

$$u_I(x_j) = u(x_j), \quad j = 0, \cdots, N.$$

It is evident that $u_I(x) = \sum_{j=0}^{N} u(x_j)\phi_j(x)$.

3.1.1 Interpolation Error bounded by $L^\infty-$norm

Theorem 3.4.

$$\|u - u_I\|_\infty \leqslant \frac{h^2}{8} \max_{x\in\bar{I}} |u''(x)|.$$

4 $P2-$ FEM

5 Implementation in General Framework

5.1 Target Problem

5.2 Finite Element Spaces

5.3 Finite Element Discretization

5.4 Boundary Treatment

5.5 Finite Element Method

References

[Evans (2010)] Evans L C. Partial differential equations[M]. American Mathematical Society, Second Edition, 2010.